Election Result Prediction based on Twitter Data Analysis

Krešimir Bačurin, Ivan Svalina, Toni Rončević

Abstract

Following article aims to demonstrate ability to predict results of elections based of posts on social networks. While the first preelection poll was conducted back in 1824 and had only 335 voters, today number of votes within poles is in the range of hundred thousands. Not only is the number of voters much higher, but the funds and science behind it are exponentially higher. The article will show how it is possible to somewhat accurately predict a political race using only data from widely available posts on social media. Dataset that is going to be used within the article is extracted from Twitter, quite possibly the most popular social media when it comes to expressing political opinions. Development will be discussed in the following chapters.

Keywords – NLP, NLTK, Twitter, Python

Paper Type – Research paper.

1. Introduction

Elections are one of the main forms of democracy. Citizens decide which political body or person will represent them for a certain number of years. The results obtained by election prediction can be used by citizens to change their choice or by political body to change their campaign.

Nowadays, people use social networks more and more. They use them to show the activities they carry out on a daily basis, but often express their political leaning. The purpose of this project is a quick and easy evaluation on who could be the potential winner of the election based on people's tweets. It greatly shortens the long, laborious and expensive process of conducting surveys, which are not necessarily a worse option because they can be more precise, but they are more difficult to conduct.

Sentiment analysis from Twitter replies would be a pretty hefty job for a single human to do. The reason is the amount of data collected from mining from Twitter replies. That’s where NLP sentiment analysis becomes very helpful. It turns jobs that would usually take a few hours into just a few seconds. This study aims to explore methods of sentiment analysis and to analyze the data that comes from said sentiment analysis. For sentiment analysis we are using Textblob, which is NLTK based Python library for text processing. This approach is very helpful in handling big datasets and it can analyze any sentence or expression in multiple languages. That is why it is used by many in the world for data collection and analysis.

In this paper we are going to view a potential US presidential elections that would take place in January of 2023. Presidential election in the United States have a characteristic that there are two parties that are competing for the office, one being Republican and the other being Democrat. On the Democrats side we made Joe Biden, the current president, as a candidate and on the Republican side we choose Kevin McCarthy, the Speaker of the House of representatives. Previously, we wanted to choose Donald Trump as a candidate, but his account was suspended two years ago and so he doesn’t have a threshold of 10000 replies in the last week on his Twitter account.

2. Related Work

Usage of social networks has skyrocketed in the past few years, one of major reasons why was the pandemic of COVID-19, and one that is used the most in expressing of opinions and emotions about the situation in geopolitical scene is Twitter[1][2]. There is publicly available Twitter API that gives us access to a large amount of tweets for our dataset, as it is described in detail in the following article [3][4].

Access to such a large volume of data opens a possibility of conducting a variety of analysis. All posts also open access to users profile from which we can extract data to conduct research on specific demographic groups, such as comparing opinions between men and women on football [5].

Predicting results of elections has been done almost as long as elections have been conducted[6], however using one of the more recently measured stat when it comes to statistics is happiness within a country [7]. We can try and predict correlation between happiness and willingness to get involved into politics. Results show that happier people tend to not get involved with politics, but when they do they tend to vote for incumbent parties[8].

Using Twitter as a source of data to predict an outcome of elections has been done through many means, and by using it as a source come certain threats that can affect the outcome of the study. Some of the threats that affect the outcome are : sarcastic tweets [9], multiple tweets from same profile entering the study.

Twitter can’t be the only online platform used to gather large amounts of data to conduct research, even platforms that were intended for gaming can be used as a ground to conduct studies and research, such as xbox Gaming on which they conducted a study to try and predict the elections in 2012 [10].

3. Methods and results

3.1 Preprocessing and Dataset

We scrapped 20000 tweet replies from the Twitter Developer API, 10000 for Kevin McCarthy and Joe Biden respectively, and we analyzed their sentiments toward either presidential candidate. We were using Python and Tweepy to scrap these Tweets. Tweepy is an open source Python package that enables access to Twitter API with Python. Tweepy's method api.search_tweets was quiet useful to us. It was searching for latest tweet replies to each presidential candidate and returned a collection of relevant Tweets matching result_type. In our case we put recent tag which enables us to scrap Tweets from last 7 days, out of which we filtered last 10000 tweets. Other tags are mixed(which returns both popular and negative comments in real time) and popular(which scraps only popular results in the response). Regarding other parameters in the method, q is a search query string of 500 characters maximum and timeout is a maximum amount of time to wait for twitter results. Also it is important to say that when running the code the maximum amount of data that can be scrapped from Twitter API is 900 requests over a 15 minute interval. When the program sees that it has sent too 900 requests it goes into a sleep mode for 15 minutes and then it sends 900 requests again does it again and again.

Replies had different attributes to it, such as: ID of a user who wrote the reply, the reply itself, profile text color etc. After we scrapped Twitter replies we wrote replies for McCarthy and Biden into .csv files. Tweets were written in .csv files as a dictionary with username as a key and tweet reply text as a reply

3.2 Text processing-TextBlob Sentiment analysis

Our focus for the analysis were the words in replies that express some kind of sentiment. We were looking for sentiments of polarity and subjectivity from a reply. For this purpose we were using Textblob Python library which provides a simple API for diving into common natural language processing(NLP) tasks, such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation and more[11]. It was built on the shoulders of NLTK and Patterns. NLTK or Natural Language Toolkit is a go-to API for NLP processing with Python and also it is a powerful tool to preprocess text data for analysis. First step when working with Textblob is tokenization where Textblob divides a body of text into words or sentences. After that Textblob scraps stop words(such as and, or, they etc.) in a text and removes them from the list of words from the previously tokenized text. After that comes stemming, where Textblob 'fluff' words are removed from a word and grouped together with it's stem form. Example of this is words such as play, playing, played, playful where stemmer stems them into just "play". Step after stemming is Tagging parts of speech where each already filtered word is tagged with a corresponding parts of speech identifier as tuples. For example ('Walk','VB') which means walk verb.[12] Also Textblob can be used in other functions such as converting text to a singular and plural, noun phrase extraction, spelling correction and so on.

After we have our text processed, we are using words from the text to have our sentiment analysis. Sentiment analysis is done with Textblob library where it has words and expressions stored in a Textblog objects rated for subjectivity and polarity. The objects is stored as a tuple of polarity and subjectivity. Polarity is valued between -1 and 1, -1 being most negative, 1 being most positive and 0 being neutral. Subjectivity is valued from 0 to 1 and it quantifies the amount of personal opinion and factual information in the text. For example in this case ***Sentiment(polarity=0.45, subjectivity=0.78)*** where it is seen that a reply is positive and very subjective.[13]

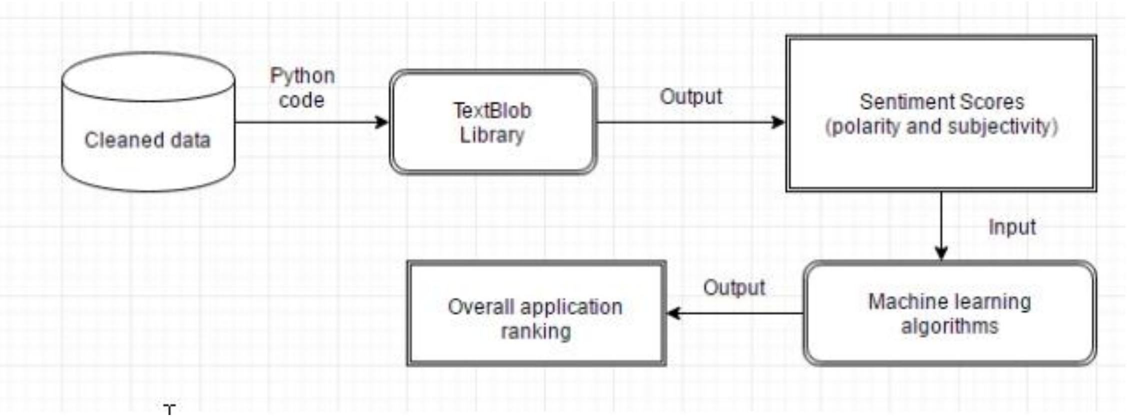


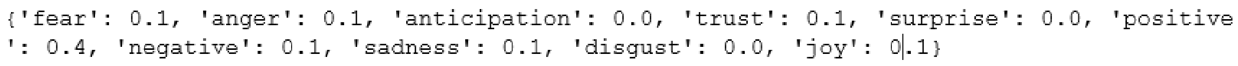
Figure 3.2.1. Textblob NLP

3.3 Text processing-Vader sentiment analysis

Vader is a python library that is specfically attuned to sentiments expressed by social media. Vader uses a combination of a sentiment lexicon which is a list of lexical features which are labeled according to their semantic orientation as either positive or negative. Vader generates a dictionary with 4 keys : neg, neu, pos and compound. Sum of all values is always 1.Compound is calculated as the sum of valence score of each word in the lexicon and detemnies the degree of sentiment. Vader also has the capability to tell us how negative or positive a certian post is which gives a possiblity to rank them, however for the use of this project we haven't used that feature.

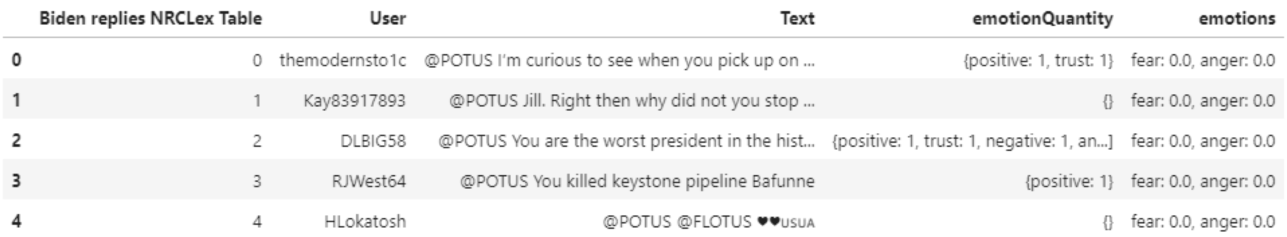
3.4 Text processing-Emotion analysis

In the second part of text processing we were analyzing emotion from tweet replies directed both to McCarthy and to Biden. Library that was used for emotion analysis is NRCLex. It is an MIT-approved PyPI project made by Mark M. Bailey which predicts sentiments and emotion on any text. Library contains approximately 27000 words and is base on the National Research Council Canada(NRC) affect lexicon and the NLTK library’s WordNet synonym sets. Emotional affects which it analysis are fear, anger, anticipation, trust, surprise, positive, negative, sadness, disgust and joy[14]. From these emotional affects it can make various derivatives and calculate different effects using gathered data. Of those various methods in this text we are using raw_emotion_scores and affect_frequencies. Raw emotion scores are counting the number of times each emotion affect is seen in any reply and on the other hand Affect frequency measures the strength and frequency of each emotion affect. The frequency gives as information as to which emotional affect weighs more than other emotional affects. It’s sum is always 1 and values must be arranged by their weight in the reply. From these two methods we have collected all the data that is presented after emotion analysis. Example of data collected is shown in tables bellow



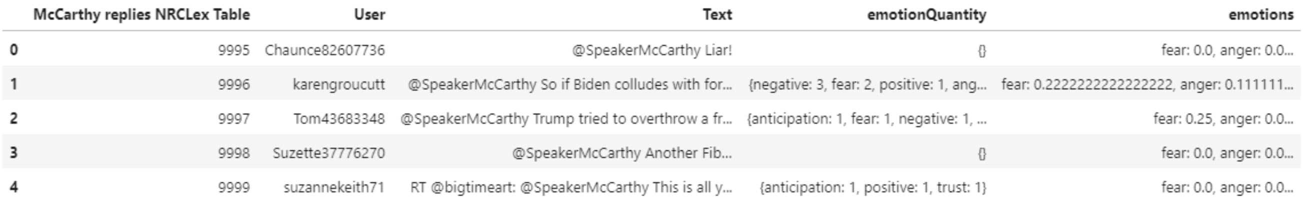
```
{'fear': 0.1, 'anger': 0.1, 'anticipation': 0.0, 'trust': 0.1, 'surprise': 0.0, 'positive
': 0.4, 'negative': 0.1, 'sadness': 0.1, 'disgust': 0.0, 'joy': 0.1}
```

Figure 3.4.1 Emotion affect frequency example



	Biden replies NRCLex Table	User	Text	emotionQuantity	emotions
0	0	themodernsto1c	@POTUS I'm curious to see when you pick up on ...	{positive: 1, trust: 1}	fear: 0.0, anger: 0.0
1	1	Kay83917893	@POTUS Jill. Right then why did not you stop ...	{}	fear: 0.0, anger: 0.0
2	2	DLBIG58	@POTUS You are the worst president in the hist...	{positive: 1, trust: 1, negative: 1, an...]	fear: 0.0, anger: 0.0
3	3	RJWest64	@POTUS You killed keystone pipeline Bafunne	{positive: 1}	fear: 0.0, anger: 0.0
4	4	HLokatosh	@POTUS @FLOTUS ♥♥USUA	{}	fear: 0.0, anger: 0.0

Photo 3.4.2. Biden replies NRCLex Table example



	McCarthy replies NRCLex Table	User	Text	emotionQuantity	emotions
0	9995	Chaunce82607736	@SpeakerMcCarthy Liar!	{}	fear: 0.0, anger: 0.0...
1	9996	karengroucutt	@SpeakerMcCarthy So if Biden colludes with for...	{negative: 3, fear: 2, positive: 1, ang...	fear: 0.2222222222222222, anger: 0.111111...
2	9997	Tom43683348	@SpeakerMcCarthy Trump tried to overthrow a fr...	{anticipation: 1, fear: 1, negative: 1, ...	fear: 0.25, anger: 0.0...
3	9998	Suzette37776270	@SpeakerMcCarthy Another Fib...	{}	fear: 0.0, anger: 0.0...
4	9999	suzannekeith71	RT @bigtimeart: @SpeakerMcCarthy This is all y...	{anticipation: 1, positive: 1, trust: 1}	fear: 0.0, anger: 0.0...

Figure 3.4.3 McCarthy replies NRCLex Table example

3.5 Topic Analysis

Topic modeling is an unsupervised machine learning technique that's capable of scanning a set of documents, detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions that best characterize a set of documents. Data analysis of social media posts, emails, chats, open-ended survey responses, and more, is not an easy task, and less so when delegated to humans alone. That's why many are excited about the implications artificial intelligence could have on their day-to-day tasks, as well as on businesses as a whole. AI-powered text analysis uses a wide variety of methods or algorithms to process language naturally, one of which is topic analysis – used to automatically detect topics from texts.

The Twitter text data is dirty. Before we get into the modeling process, the data was cleared first. Therefore, we can get high-quality information from it. To clean the text, we used libraries like "NLTK" and "re". That library filters words, mentions, hashtags, links, and many more. After we have cleaned the data, we did the topic modeling process. For the modeling process, we used the BERTopic library. The BERTopic model will fit and transform the tweets for generating topics based on the tweets. From there on we could retrieve topics that exist on the dataset.

3.6 Analytics

After we determined polarity and subjectivity of certain replies, we filtered replies according to polarity into negative(from -1 to 0), neutral(0) and positive(from 0 to 1). Then we counted each type of replies and used it for further analysis. In regards to emotion affect analysis, our goal was to examine the affective context in which certain emotions rise in correlation with other emotions and to present statistical data in regards to emotion. Also we are going are comparing sentiment analysis with emotional analysis in regards to classification as positive or negative. And lastly we are going to examine emotion correlation of emotional affects of negative with supposed negative affects such as fear and anger and positive with supposedly positive traits such as trust and joy

4. Results and Discussion

4.1 TextBlob Sentiment analysis

From graphs below we can see that there are many cases where the polarity is 0. This is caused by tweets that do not have any text, or contain only links and hashtags. For this reason, these tweets will be discarded, because in order to evaluate who has a greater chance of winning the presidential elections, we need positive and negative tweets, not neutral ones from which it is difficult to conclude anything. However, we can notice that McCarthy has more neutral tweets than Biden, and the most likely reason for this is that McCarthy's function is less in the attention of the media and also a lot less responsible than the function of the president of the United States. Also Kevin McCarthy just came into office, so there are a lot of welcoming messages

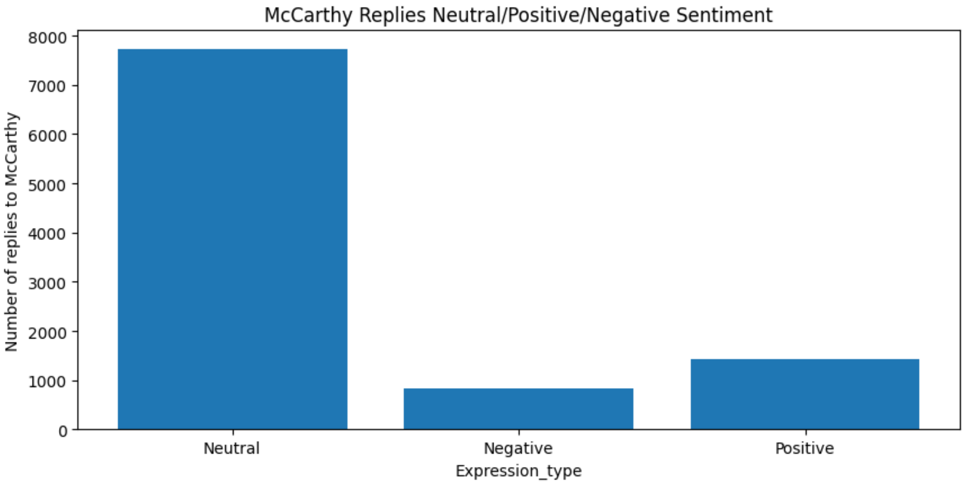


Figure 4.1.1. Replies to McCarthy sentiments

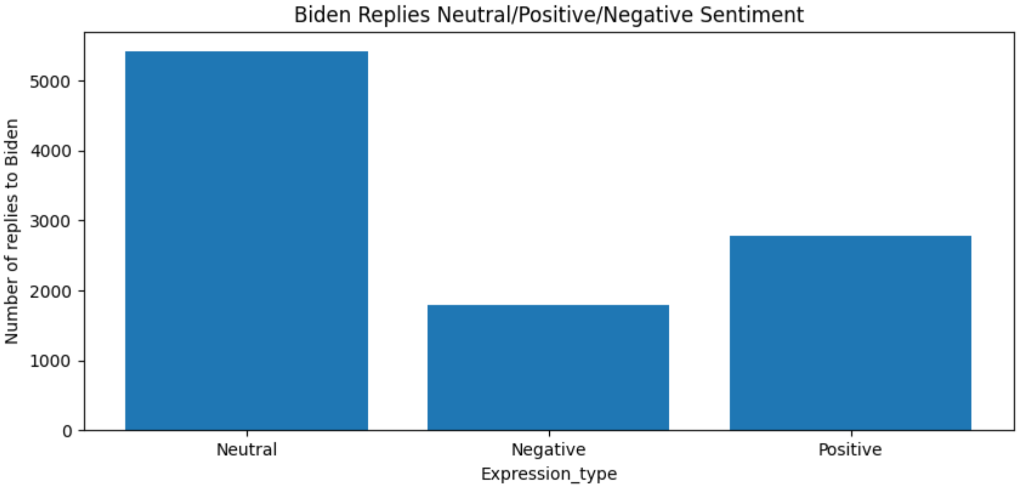


FIgure 4.1.2. Replies to Biden sentiments

From the attached below, it is clearly visible that many more negative comments were directed at Biden than at McCarthy, but at the same time, McCarthy had fewer positive comments than Biden. One possible cause is that Biden is the current president and is under the watchful eye of the nation. People will write a lot of negative comments, but there will also be positive ones. These results could change just before the election when both would be equally under the public limelight and therefore have similar number of tweets directed towards them.

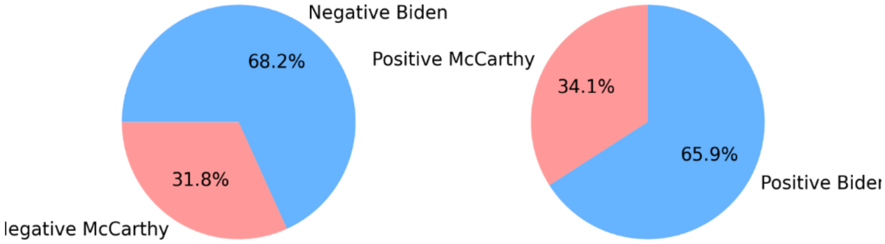
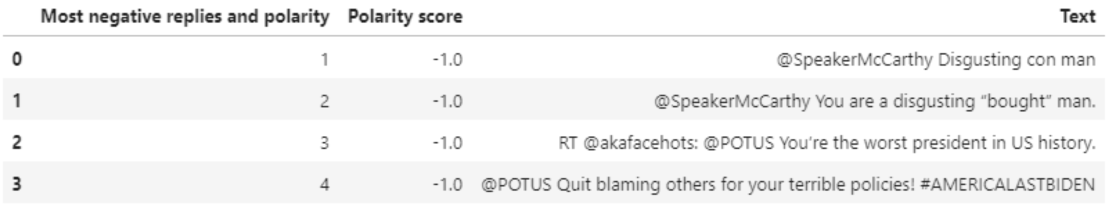
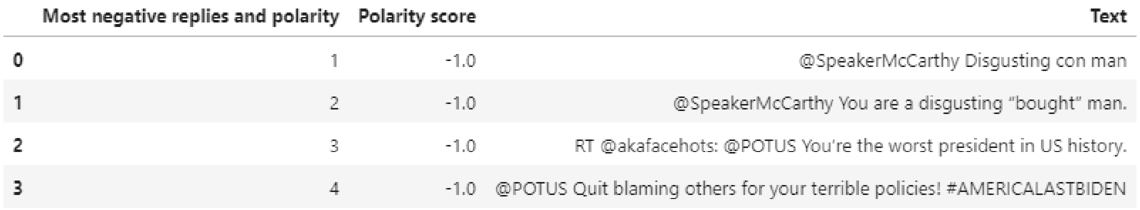


Figure 4.1.3. Pie chart negative vs. positive reply



	Most negative replies and polarity	Polarity score	Text
0	1	-1.0	@SpeakerMcCarthy Disgusting con man
1	2	-1.0	@SpeakerMcCarthy You are a disgusting “bought” man.
2	3	-1.0	RT @akafacehots: @POTUS You're the worst president in US history.
3	4	-1.0	@POTUS Quit blaming others for your terrible policies! #AMERICALASTBIDEN

Figure 4.1.4. The most positive replies to both candidates



	Most negative replies and polarity	Polarity score	Text
0	1	-1.0	@SpeakerMcCarthy Disgusting con man
1	2	-1.0	@SpeakerMcCarthy You are a disgusting “bought” man.
2	3	-1.0	RT @akafacehots: @POTUS You're the worst president in US history.
3	4	-1.0	@POTUS Quit blaming others for your terrible policies! #AMERICALASTBIDEN

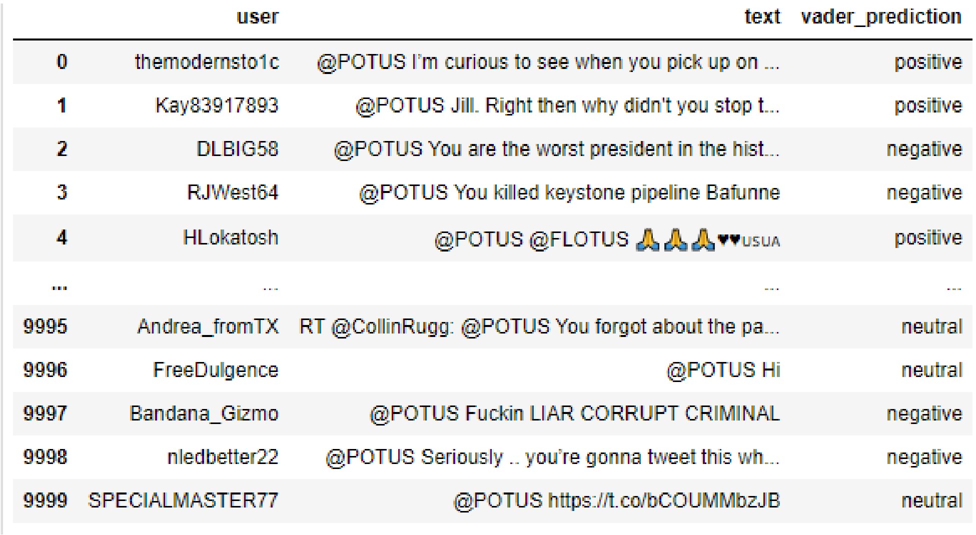
Figure 4.1.5. The most negative replies to both candidates

Above we can see a lot of the most positive and negative tweets directed at Biden and McCarthy. It is interesting to analyze this because something can be concluded. When we take a closer look at the most positive tweets directed at Biden, we can specifically notice these two sentences: "*They'll have WiFi but no food. Awesome idea!*", "*The best thing you can do for this country is resign… please*". The program will understand these sentences as positive comments with the most positive polarity of 1.0, however, when we humans read these sentences, it will immediately be clear to us that it is sarcasm. However, this is not necessarily a bug in the program because we humans are also sometimes incapable of understanding sarcasm.

It should also be considered that extremely negative comments can often have a greater and stronger impact than positive ones. Often besmirching of one candidate can make undecisive candidates make a decision more easily, and can also make voters who, when they see negative comments, change their decision in favor of the opposing candidate.

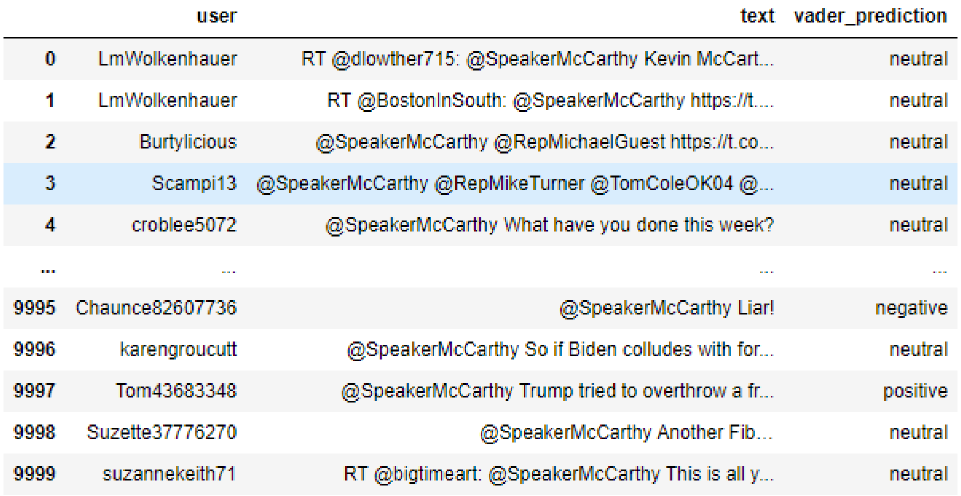
4.2 Vader sentiment analysis

After running our dataframe through vader results were as follows. Biden had a total of 2523 positive tweets, 3388 negative tweets and a total of 4089 tweets that vader considered neutral. McCarthy had 1748 positive tweets, 1634 negative tweets and a total of 6618 neutral tweets. Following pictures show Biden's and McCarthy's datasets.



	user	text	vader_prediction
0	themodernsto1c	@POTUS I'm curious to see when you pick up on ...	positive
1	Kay83917893	@POTUS Jill. Right then why didn't you stop t...	positive
2	DLBIG58	@POTUS You are the worst president in the hist...	negative
3	RJWest64	@POTUS You killed keystone pipeline Bafunne	negative
4	HLokatosh	@POTUS @FLOTUS 🙏🙏🙏♥♥USUA	positive
...	...	...	...
9995	Andrea_fromTX	RT @CollinRugg: @POTUS You forgot about the pa...	neutral
9996	FreeDulgence	@POTUS Hi	neutral
9997	Bandana_Gizmo	@POTUS Fuckin LIAR CORRUPT CRIMINAL	negative
9998	nledbetter22	@POTUS Seriously .. you're gonna tweet this wh...	negative
9999	SPECIALMASTER77	@POTUS https://t.co/bCOUMMbzJB	neutral

Figure 4.2.1 Vader analysis Biden

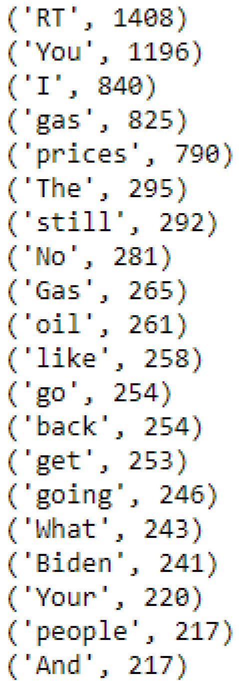


	user	text	vader_prediction
0	LmWolkenhauer	RT @dlowther715: @SpeakerMcCarthy Kevin McCart...	neutral
1	LmWolkenhauer	RT @BostonInSouth: @SpeakerMcCarthy https://t....	neutral
2	Burtylicious	@SpeakerMcCarthy @RepMichaelGuest https://t.co...	neutral
3	Scampi13	@SpeakerMcCarthy @RepMikeTurner @TomColeOK04 @...	neutral
4	croblee5072	@SpeakerMcCarthy What have you done this week?	neutral
...	...	...	...
9995	Chaunce82607736	@SpeakerMcCarthy Liar!	negative
9996	karengroucutt	@SpeakerMcCarthy So if Biden colludes with for...	neutral
9997	Tom43683348	@SpeakerMcCarthy Trump tried to overthrow a fr...	positive
9998	Suzette37776270	@SpeakerMcCarthy Another Fib...	neutral
9999	suzannekeith71	RT @bigtimeart: @SpeakerMcCarthy This is all y...	neutral

FIgure 4.2.2 Vader analysis Mccarthy

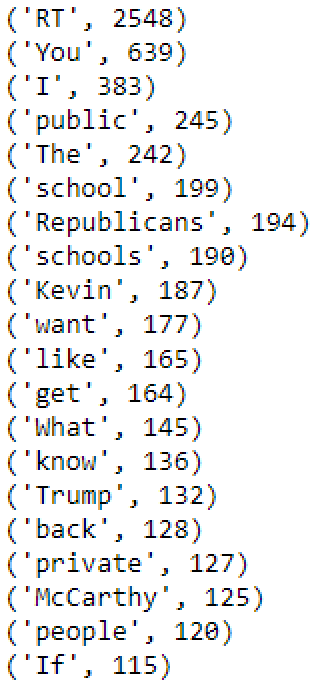
Analysis shows that Biden has slightly more negative tweets targeted towards him which is to be expected out of President who is currently active, while for McCarthy has a large majority of neutral tweets.

In order to count the most commonly used words we used a built in python library to get rid of all stopwords (such as the, is, are...), because obviously those would be the most commonly used ones in tweets, and after that we got rid of all twitter tags and mentions. Following pictures show most commonly used words in tweets for Biden and Mccarthy in that order in form of touples.



```
('RT', 1408)
('You', 1196)
('I', 840)
('gas', 825)
('prices', 790)
('The', 295)
('still', 292)
('No', 281)
('Gas', 265)
('oil', 261)
('like', 258)
('go', 254)
('back', 254)
('get', 253)
('going', 246)
('What', 243)
('Biden', 241)
('Your', 220)
('people', 217)
('And', 217)
```

Figure 4.2.3 Most used words Biden



```
('RT', 2548)
('You', 639)
('I', 383)
('public', 245)
('The', 242)
('school', 199)
('Republicans', 194)
('schools', 190)
('Kevin', 187)
('want', 177)
('like', 165)
('get', 164)
('What', 145)
('know', 136)
('Trump', 132)
('back', 128)
('private', 127)
('McCarthy', 125)
('people', 120)
('If', 115)
```

Figure 4.2.4 Most used words Mccarthy

Out of the most common words the majority of them doesn't reflect any kind of sentiment. There are a loto f words that are common in political discourse such as: like, people, private, school… Other words that are not reflecting any sentiment are pronouns, names, verbs and etc. It's very interesting that the most common words are RT, you and I. RT is to be expected to be the most common word on Twitter, because it referes to retweeting. In regards to words that reflect sentiment it is interesting that the topic of gas and oil are most common in tweets directed at Biden, where towards McCarthy there isn't an obvious topic. The reason for this are high energy prices that are causing a lot of dissatasfaction towards Biden. On the other hand the most common words that reflect sentiment for McCarthy are school and Trump. School is shown in these stats because of McCarthy's engagement in public school reform. This can be seen both positive because any help to schools is quiet popular. The second McCarthy's keyword is Trump, because of his dominance in the Republican spectrum. It's quiet interesting to see that even after two years of being blocked on Twitter, there are still a huge amount of replies that mention Trump. Even more interesting dana is that keyword Trump is more mentioned than McCarthy in replies to McCarthy. This is quiet an interesting stat that would have effect in presidential election between McCarthy and Biden. It is surely a stat that goes against McCarthy. This explains a much more neutral attitude towards McCarthy in tweet replies than to Biden. Just from the sheer difference in graphs it is obvious how Biden has words that are more represented than McCarthy, whose graph is relatively flat other than a single word.

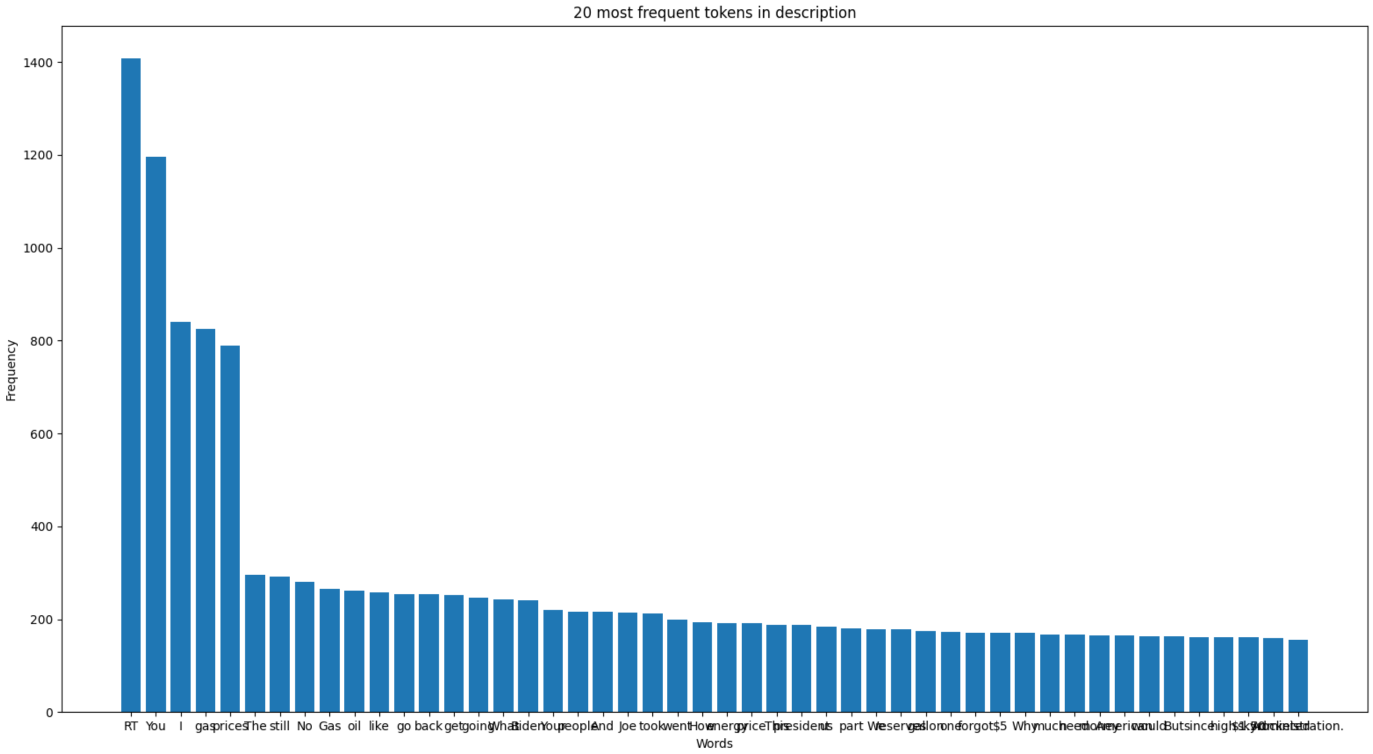


Figure 4.2.5 Most frequent words Biden chart

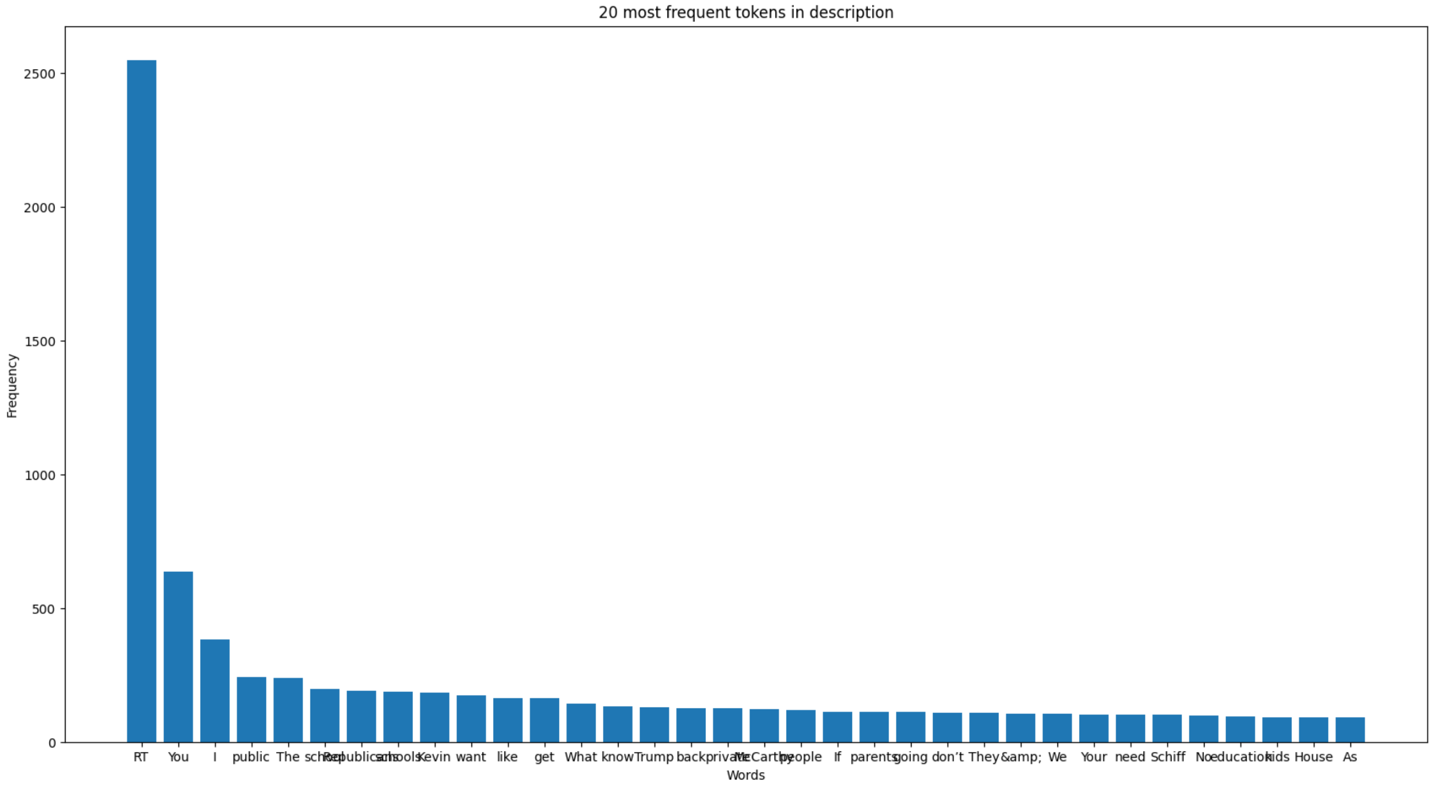
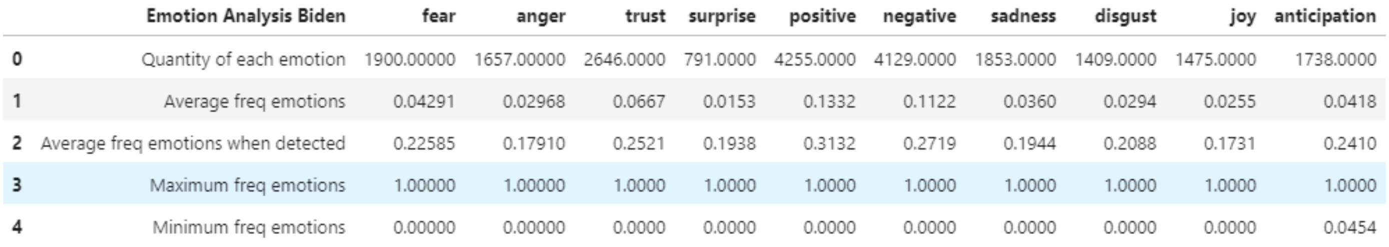


Figure 4.2.6 Most frequent words Mccarthy chart

4.3 Emotion analysis- Results and discussion

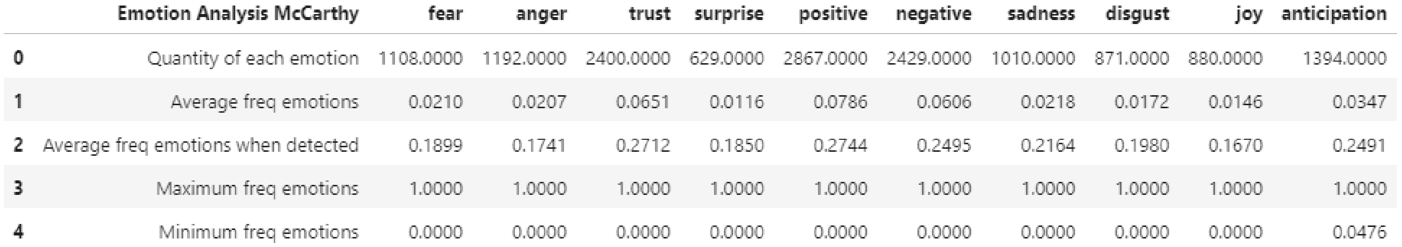
Firstly we have generated data representing emotion affect occurrence and emotion affect frequency in each tweet reply. After that we have calculated different metrics for each emotion. Quantity of each emotion was counted by counting raw emotion scores from each reply. Average affect frequency of emotions is an average affect frequency of each emotional affect even when it is 0. As we can see all of emotional affect frequencies have pretty low values. The reason for it is that a lot of the times affect frequencies are 0. This correlates with a high number of neutral comments detected in sentiment analysis with TextBlob. That’s why we have calculated average emotional affect frequency when the affect frequency is higher than 0. We can see that the amount is higher than the former, but also that all of the numbers calculated are lower than 0.35. As we can see the emotional affect frequencies, even when they are detected are not excessive as one would think. It is because certain emotional affects come in pair with others: negative with anger, fear, sadness and disgust and positive with joy, surprise and trust especially. And lastly we have calculated maximum and minimum emotional affect frequency. For each emotional affect maximum affect frequency is 1 and minimum emotional affect frequency is 0 respectively, except in the case of anticipation. In the case of anticipation the minimal value is a bit bigger than 0, and is between 0.045 and 0.5 for both candidates. The reason for this probably lies in the fact that politics is quiet a nerve wracking, tense and very emotional topic.

In comparison with sentiment analysis which was made using TextBlob it can be seen that positive and negative affects are more balanced with NRCLex for both candidates. In TextBlob relation between positive replies are 3/2. But when we take into account other positive and negative emotional affects such as trust and fear we can see that their ratio is close to ratio calculated in TextBlob. So we conclude that results from these two libraries correlate. Also it is important to notice that NRCLex is unable to recognize sarcasm like TextBlob. For example tweet: *kbell43,@POTUS So the administration is only responsible for gas prices when they go down?* has 0.5 positive and 0.5 trust emotional affect frequency.



	Emotion Analysis Biden	fear	anger	trust	surprise	positive	negative	sadness	disgust	joy	anticipation
0	Quantity of each emotion	1900.00000	1657.00000	2646.0000	791.0000	4255.0000	4129.0000	1853.0000	1409.0000	1475.0000	1738.0000
1	Average freq emotions	0.04291	0.02968	0.0667	0.0153	0.1332	0.1122	0.0360	0.0294	0.0255	0.0418
2	Average freq emotions when detected	0.22585	0.17910	0.2521	0.1938	0.3132	0.2719	0.1944	0.2088	0.1731	0.2410
3	Maximum freq emotions	1.00000	1.00000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4	Minimum freq emotions	0.00000	0.00000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0454

Figure 4.3.1 Biden emotional analysis



	Emotion Analysis McCarthy	fear	anger	trust	surprise	positive	negative	sadness	disgust	joy	anticipation
0	Quantity of each emotion	1108.0000	1192.0000	2400.0000	629.0000	2867.0000	2429.0000	1010.0000	871.0000	880.0000	1394.0000
1	Average freq emotions	0.0210	0.0207	0.0651	0.0116	0.0786	0.0606	0.0218	0.0172	0.0146	0.0347
2	Average freq emotions when detected	0.1899	0.1741	0.2712	0.1850	0.2744	0.2495	0.2164	0.1980	0.1670	0.2491
3	Maximum freq emotions	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4	Minimum freq emotions	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0476

Figure 4.3.2 McCarthy emotional analysis

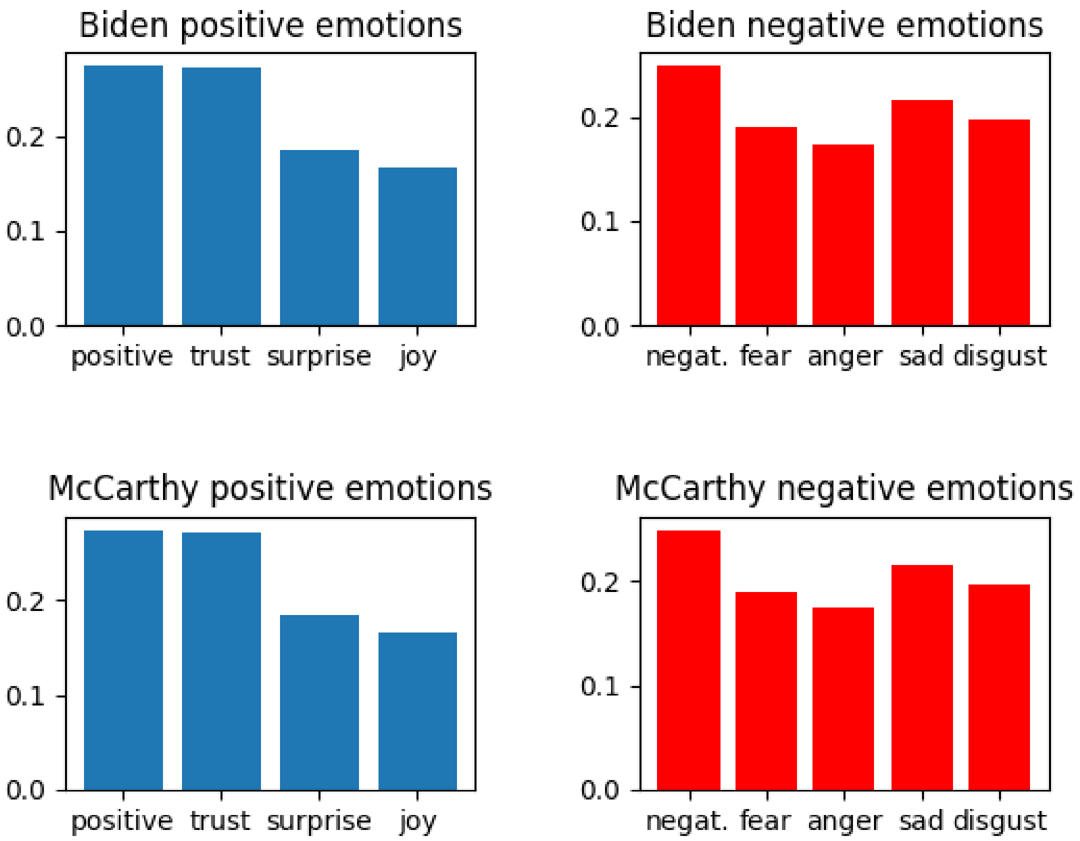


Figure 4.3.3 Positive/negative emotions ratio in regards to avg affect frequency when detected

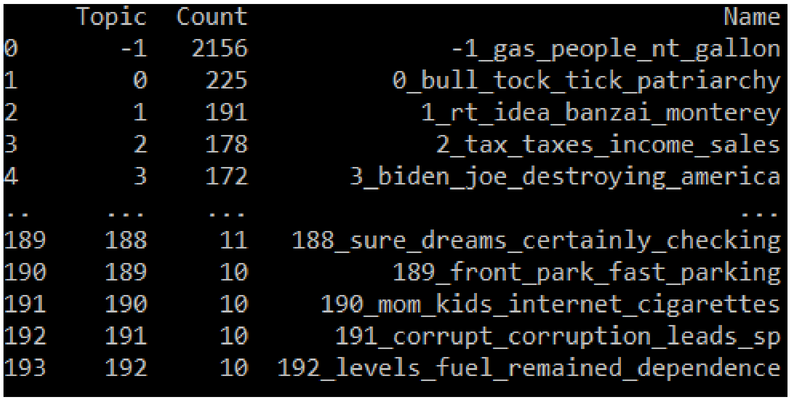
From analyzing tweet replies we have noticed that certain emotions come in pair. So are positive, trust, surprise and joy and negative, fear, anger , sadness and disgusting coming in pairs. For purposes of our paper we are going to name them 'positive' and ' negative'. Both candidates have similar ratios between different 'positive' and 'negative' emotions and we can analyze both together. Positive and trust have almost the same affect frequency and other 'positive' emotions like surprise and joy are a less detected then former two. This is probably because of the political topics, where trust is quiet important. On the other hand in 'negative' emotions negative is the one most detected. After that comes sadness and a bit bellow sadness come other three emotions. It's easy to see that the ratio between different 'negative' emotions is more balanced than in 'positive' emotions. It is quiet known that Twitter generates a lot of negative and hate comments and so it is visible in these reply sections

When comparing the tables for each candidate we can see that Biden's tweets create much more emotion around them than McCarthy's. If we look carefully at the data collected we can see that in regards to emotions correlated with 'negative' and some 'positive'(positive, joy) there is almost a twice as much quantity and average affect frequency of each emotion detected in Biden's than in McCarthy's replies. On the other hand when it comes to emotional affect frequencies trust and surprise McCarthy has almost equal average affect frequency and quantity as Biden. This is especially seen in the average frequency affect emotions when detected where difference between them is decreased and in emotions trust and sadness McCarthy is higher than Biden's

The reason for this kind of relation in results between Biden and McCarthy is that Biden's position has much more responsibility and therefore it has stronger emotions connected to him unlike McCarthy. McCarthy's results are connected to him just coming to office and consequently it leads to higher trust and sadness than average. Trust because he probably gets a lot of congratulatory replies and sadness because users that haven't voted for him are sad because he got into office.

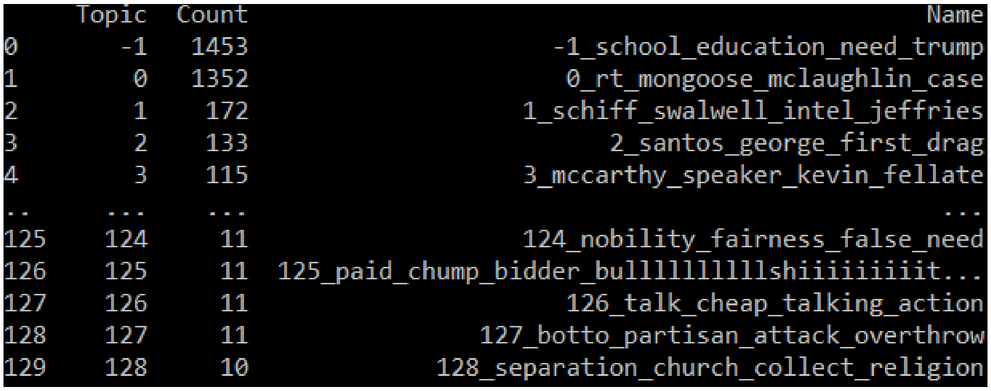
4.4 Topics

Following pictures show Biden's and McCarthy's topic count. There are 193 topics in Biden dataset and 129 topics in Mccarthy dataset. The -1 topic doesn’t include because those tweets don’t have a significant meaning. The number of repetitions of a certain topic is shown, from the most frequent to the least frequent one.



```
     Topic  Count                               Name
0       -1   2156             -1_gas_people_nt_gallon
1        0    225           0_bull_tock_tick_patriarchy
2        1    191            1_rt_idea_banzai_monterey
3        2    178             2_tax_taxes_income_sales
4        3    172        3_biden_joe_destroying_america
..     ...    ...                                  ...
189    188     11    188_sure_dreams_certainly_checking
190    189     10           189_front_park_fast_parking
191    190     10       190_mom_kids_internet_cigarettes
192    191     10        191_corrupt_corruption_leads_sp
193    192     10  192_levels_fuel_remained_dependence
```

Figure 4.4.1 Biden topic count



```
     Topic  Count                                               Name
0       -1   1453                    -1_school_education_need_trump
1        0   1352                        0_rt_mongoose_mclaughlin_case
2        1    172                     1_schiff_swalwell_intel_jeffries
3        2    133                          2_santos_george_first_drag
4        3    115                     3_mccarthy_speaker_kevin_fellate
..     ...    ...                                                ...
125    124     11                     124_nobility_fairness_false_need
126    125     11  125_paid_chump_bidder_bulllllllllllshiiiiiiiiit...
127    126     11                        126_talk_cheap_talking_action
128    127     11                   127_botto_partisan_attack_overthrow
129    128     10                128_separation_church_collect_religion
```

Figure 4.4.2 Mccarthy topic count

The first visualization that we can create is the distance map between topics. The intertopic distance map is a visualization of the topics in a two-dimensional space. The area of these topic circles is proportional to the amount of words that belong to each topic across the dictionary. The circles are plotted using a multidimensional scaling algorithm (converts a bunch of dimension, more than we can conceive with our human brains, to a reasonable number of dimensions, like two) based on the words they comprise, so topics that are closer together have more words in common. This is also a great tool to group topics.

We can see example of grouping on Biden’s Intertopic Distance Maps below. Topic 34 contains words such as resign, resignation, departure, enforced, deffered while topic 149 contains words such as dividing, divide, focus, divisive, unity. Overall all circles in that area are predominantly negative. This analysis can help candidates on which topics they should focus on because it can determine the result of elections. For instance if Trump had focused more on handling the covid-19 pandemic better, maybe he would be a president today.

Similar analysis is applied on Mccarthy’s Intertopic Distance Map, but it has a lot less topic groupings in it.

The slider below distance map is used to select the topic which then lights up red. If we hover over a topic, then general information is given about the topic, including the size of the topic and its corresponding words.

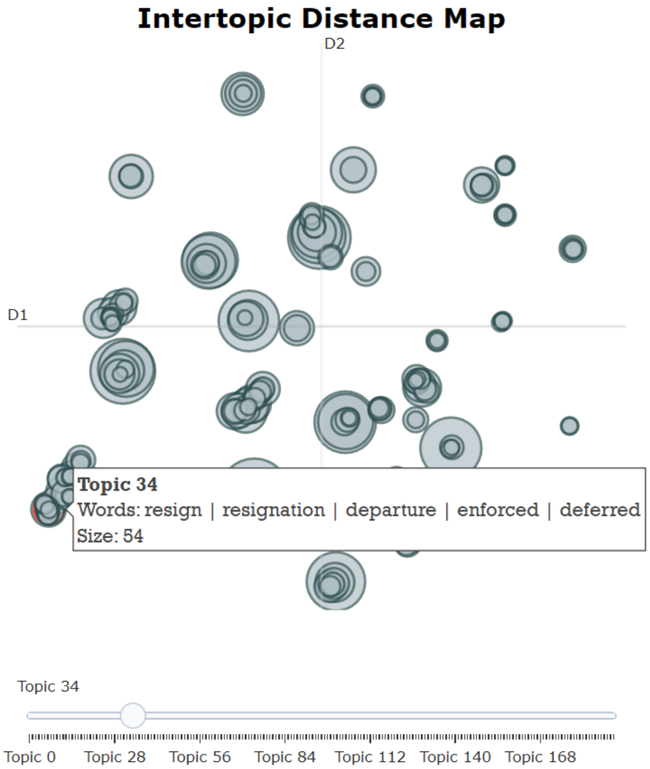


Figure 4.4.3 Biden Intertopic Distance Map 1

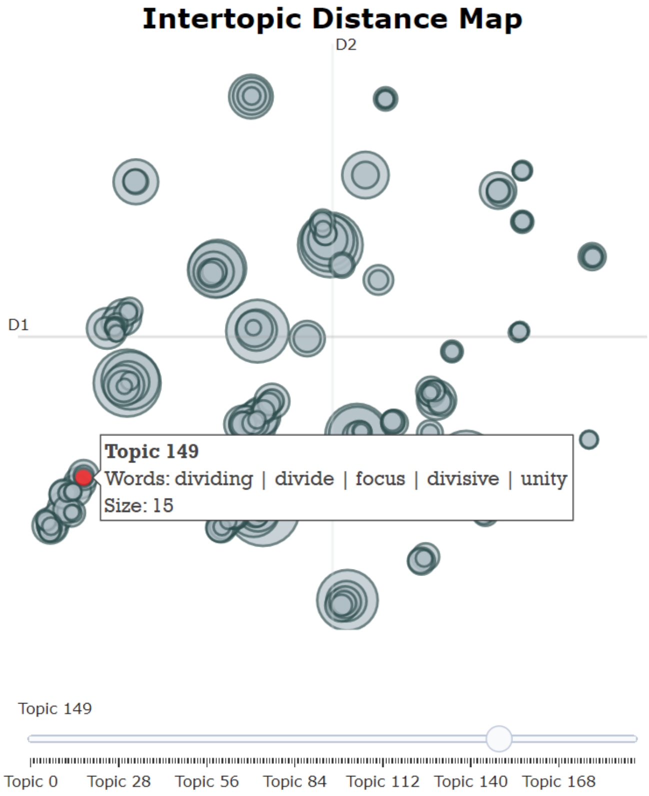


Figure 4.4.4 Biden Intertopic Distance Map 2

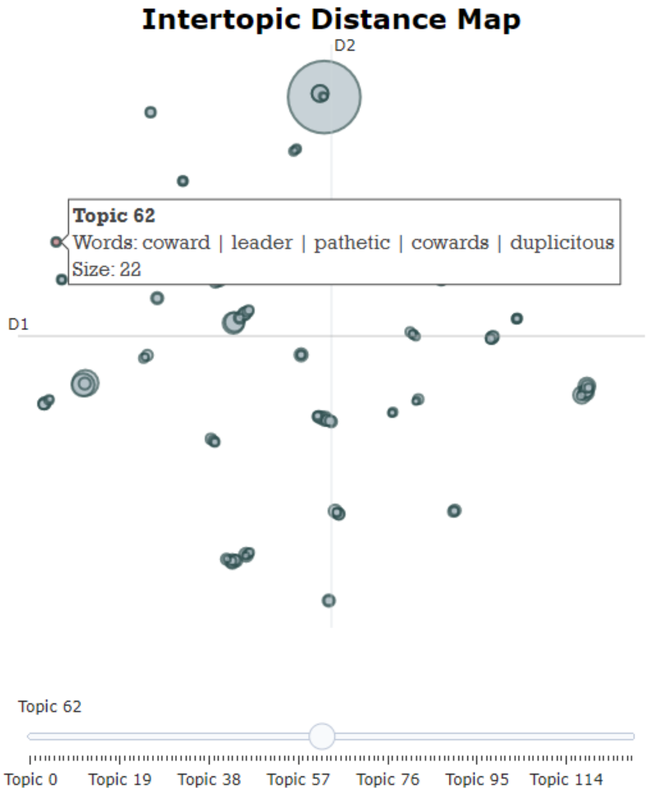


Figure 4.4.5 Mccarthy Intertopic Distance Map

The second visualization are bar charts of the most occurred words for each topic. We can visualize the selected terms for a few topics by creating bar charts out of the c-TF-IDF scores for each topic representation. TF-IDF (term frequency-inverse document frequency) is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. This is done by multiplying two metrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of documents. Moreover, we can easily compare topic representations to each other.

The bar chart by default shows most salient terms. The bars indicate the total frequency of the term across the entire corpus. Salient is a specific metric, defined at the bottom of the visualization, that can be thought of as a metric used to identify most informative or useful words for identifying topics in the entire collection of texts. Higher saliency values indicate that a word is more useful for identifying a specific topic.

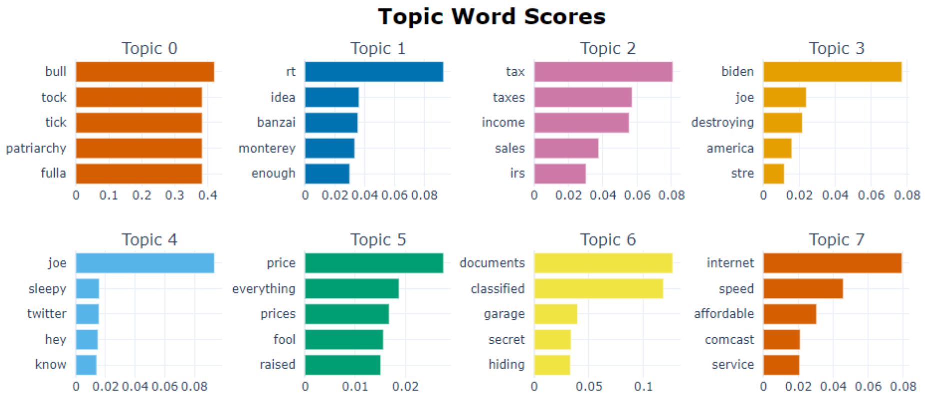


Figure 4.4.6 Biden barchart

Biden's bar chart shows the first eight most frequent topics. They, like the tweets themselves, could also be divided into positive, negative and neutral groups. Topics 3, 4, 5, and 6 would be strictly negative, while topic 2 could be positive or negative depending on people you ask. Topic 2 mentions the words "tax" and "irs" which probably refer to Biden's potential tax on high-earning companies. We can conclude topic 2 is positive based on his statements that anyone should be able to become a millionaire or a billionaire. He believes that it is wrong for America to have a tax code that results in America’s wealthiest households paying a lower tax rate than working families. In a typical year, billionaires pay an average tax rate of just 8%. In the State of the Union, he’ll call on Congress to pass his billionaire minimum tax. This minimum tax would make sure that the wealthiest Americans no longer pay a tax rate lower than teachers and firefighters.

Topics 3, 4, 5, 6 slander Biden and describe him in negative way and say that prices are rising and that Biden is destroying America. Topic 0 could potentially be positive as it mentions a bull market that signals rising stock prices and a strong economy, however based on the other words in the topic it is hard to conclude more. Topic 1 would be neutral, while Topic 7 could be considered positive because it describes an affordable Internet that probably signals a drop or stabilization of Internet service prices. As we have already mentioned in the sentiment analysis, a large number of negative themes with Biden are visible. The reason for this is that he is the current president and is under pressure from the public. People tend to write negative comments much more often than positive ones.

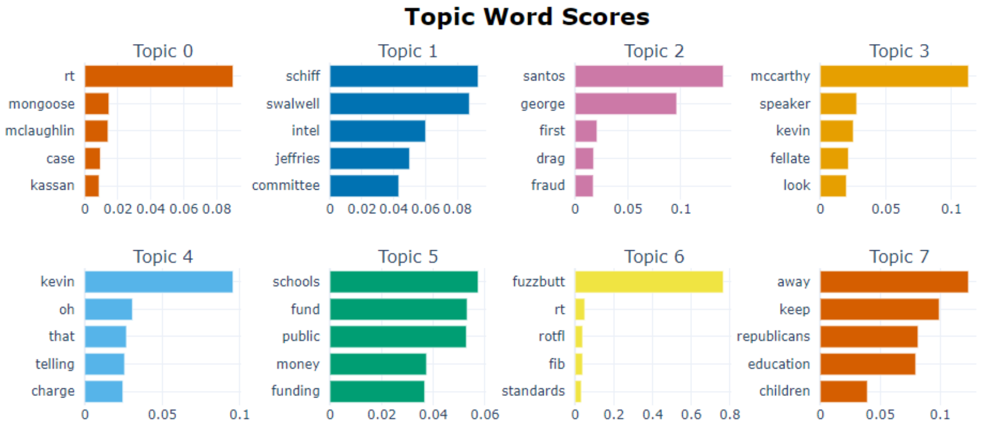


Figure 4.4.7 McCarthy barchart

On the other hand when viewing McCarthy topic word scores, one can see that they are much less connected to emotionally strong topics. In this case we can also classify the topics into positive, negative and neutral. Topic 5 would be positive, topics 2,3,6 and 7 are negative and topics 1 and 4 are viewed as neutral. Topic 0 sadly cannot be classified as any of the former. It is a topic that has RT in it and a few unconnected words and could be connected to bots. Topic 5 is classified as positive because of the increase of funding to public schools which is usually a popular move. Also it includes the increase of pay to school professors that are an important electorate. In regards to neutral topics it includes topic 1 and 4, that are classified as neutral due to them being made out of words that are connected to congress discussions and cannot be classified as either positive or negative. There are 4 topics that could be classified as negative for speaker McCarthy. Topic 2 refers to a scandal of a congressman George Santos who has supposedly lied about himself to public before election and now the information is coming out about it. Topic 3 and 6 are composed of insults directed to Kevin McCarthy and are a significant sign of negative amotion towards the speaker of the House. Topic 7 is a critic that comes from the democrat's side in regards to school legislature that McCarthy is working on. Schools are a double edged sword because you can gain votes on helping school funding, school reform etc., but people get very emotional in regards to their children so this is not unexpected

When comparing Biden and McCarthy topic analysis one can notice that Biden's topics generate much more interest and emotion than McCarthy's, but we won't say that topics of McCarthy's replies are irrelevant. We can generally see that topics that are positive to each candidate manifest in similar ways. In both cases those are things that can make people's life safer, easier and cheaper. Also in the case of Biden there is a heroic act involved so there is a big plus for him on the positive side. In case of neutral topics they are manifesting in different ways on each candidate's side. In case of Biden those are the topics of taxes and social media regulation than can go both ways or it won't go anywhere so we were confident to classify it as neutral until it is resolved whether it's positive or negative. Topics that are classified as neutral on McCarthy's replies are usually connected to the House of Congress's work and so it includes positive, neutral and negative topics. In regard to negative topics Biden's are much more serious in content. This, if grows can seriously harm Biden in future election. But it is impossible to say for now if it will grow or not. Also there are a few topics that are talking negatively about Joe Biden as a person, this can't probably harm Joe Biden's chances. But it surely has an effect. In the case of negative topics directed at McCarthy they are a two serious topics that can be quiet a determining factor in the case of election. Those topics are as serious as those directed at Biden. But topics that are personally concerning directed at McCarthy are much more direct than those directed at Biden, also they are quiet gloves-off in it's content. A thing that is also interesting to notice in the case of McCarthy is that his account is much more prone to Bot attacks. This can have an effect by interference of some foreign factor in the election.

5. Conclusion

In this project, we used twitter as a source of information, as in the previously mentioned work [3][4]. We analyzed the most positive and the most negative tweets, found most common words used, analyzed tweets in emotional sense, found topics, and it can define our decision on who we will vote for. We also encountered sarcastic comments that can affect the final result of the research, however detection and analysis of sarcastic tweets is explained in a lot more detail at [9].

The purpose of our work was to estimate who could win the next election. In regards to winning, it is known that a candidate that can rile up more of his base and have his opponents base stay home will win. When analyzing both candidate's 'positive' and 'negative' emotion with NRCLex we can see that positive are somewhat ahead. Biden's and McCarthy's ratio of 'positive' and 'negative' emotions are quiet similar. But from all of the data we can see that numbers are much stronger on the side of Biden, except for trust where McCarthy has a small lead. That means that when going to election people are probably going to vote for or against Biden. This is a double edged sword because if you have more people that think negatively of you, you are going to rile up the opposing side. But in this case Biden has a slightly higher score on the 'positive' emotion side of thing. That's why we think that in the case of the presidential elections and based on NRCLex emotion analysis Biden would be able to rile more of his base and thus win

Using Sentiment Analysis can help us decipher the mood and emotions of general public and gather insightful information. While this is an interesting way to predict elections, the results and observations collected should be taken with a grain of salt. First of all, the tweets were not collected just before the election, so we don't have the most up-to-date data, and we can be sure that not many people will write positive or hateful comments towards McCarthy because he just took office of the Speaker of the House and his potential presidential campaign hadn't even started yet. However, if we were to assess the winner now, based on TextBlob sentiment analysis, it could be Biden. Bidens comments are much more viewed as negative than positive, while McCarthy has more neutral tweets and less either positive or negative replies. If a candidate wants to win presidential elections he or she need to have bigger support of their own base and in result analysis it is visible that Biden has more positive replies which equals with more support from his own base. With that in mind, Biden might have a edge if elections were held today.

As far as the BERTopic analysis is concerned, both candidates contain a similar ratio of negative and positive topics. The difference is that Biden has a significantly larger number of tweets, and because of this, negative topics are more noticeable, but this does not necessarily mean that the public has a worse opinion of him compared to McCarthy. McCarthy does not have enough noticeable positive topics on his side that would give him a significantly greater advantage compared to the current president Biden. Based on that, we could give a greater advantage to Biden in upcoming president elections.

Based on NRCLex emotion analysis, TextBlob Sentiment analysis and BERTopic analysis Biden would be the winner in election if it were to be held today

6. Literature

[1] Mahdikhani, M (April 2022). On predicting the popularity of tweets by analizing public opinion and emotions in different stages of COVID-19. Available : https://www.sciencedirect.com/science/article/pii/S266709682100046X

[2] Pokharel, B (January 202) On Twitter sentiment analysis during Covid-19 outbreak in Nepal. Available : https://www.researchgate.net/profile/Bishwo-Prakash-Pokharel-2/publication/342228515_Twitter_Sentiment_Analysis_During_Covid-19_Outbreak_in_Nepal/links/5ef2b616458515ceb207eb07/Twitter-Sentiment-Analysis-During-Covid-19-Outbreak-in-Nepal.pdf

[3] Kumar, P and Gujjar J (April 2021) On Sentiment analysis: Textblob for decision making. Available : https://ijsret.com/wp-content/uploads/2021/03/IJSRET_V7_issue2_289.pdf

[4] Wisdom, V and Gupta, R (September 2016) On an introduction to Twitter data analysis in Python. Available : https://www.researchgate.net/profile/Vivek-Wisdom/publication/308371781_An_introduction_to_Twitter_Data_Analysis_in_Python/links/57e24cf708ae1f0b4d95b409/An-introduction-to-Twitter-Data-Analysis-in-Python.pdf

[5] Bagić Babac, M and Podobnik, V (May 2016). On a sentiment analysis of who participates, how and why, at social media sports websites. Available : https://www.fer.unizg.hr/_download/repository/How_differently_men_and_women_write_about_football[1].pdf

[6] Feddersen, T and Sandroni, A (April 2002). On a theory of participations in elections. Available : https://www.kellogg.northwestern.edu/faculty/fedderse/homepage/papers/duty4-15-02.pdf

[7] Ghosh, I (December 2018). On global happiness, which countries are the most and least happy. Available : https://www.visualcapitalist.com/measuring-global-happiness-countries/

[8] Ward, G (March 2019). On happiness and voting behaviour. Available : https://worldhappiness.report/ed/2019/happiness-and-voting-behavior/

[9] Bakliwak, A and Hughes, M (June 2013). On Sentiment analysis of political tweets : Towards an accurate classifier. Available : https://doras.dcu.ie/19962/1/foster2013.pdf

[10] Wang, W, Rothschild, D, Goel, S and Gelman, A (2014). On forecasting elections with non-representative polls. Available : https://5harad.com/papers/forecasting-with-nonrepresentative-polls.pdf

[11]Loria, S(2020.), TextBlob: Simplified Text Processing. Available: https://textblob.readthedocs.io/en/dev/

[12] Sukanta, S(November 2020.), Intro to NLTK for NLP with Python. Available: https://towardsdatascience.com/intro-to-nltk-for-nlp-with-python-87da6670dde

[13] Malik, U(July 2022.), Python for NLP: Introduction to the TextBlob Library. Available: https://stackabuse.com/python-for-nlp-introduction-to-the-textblob-library/

[14] Bailey, M.M., 2019.,NRCLex, https://pypi.org/project/NRCLex/